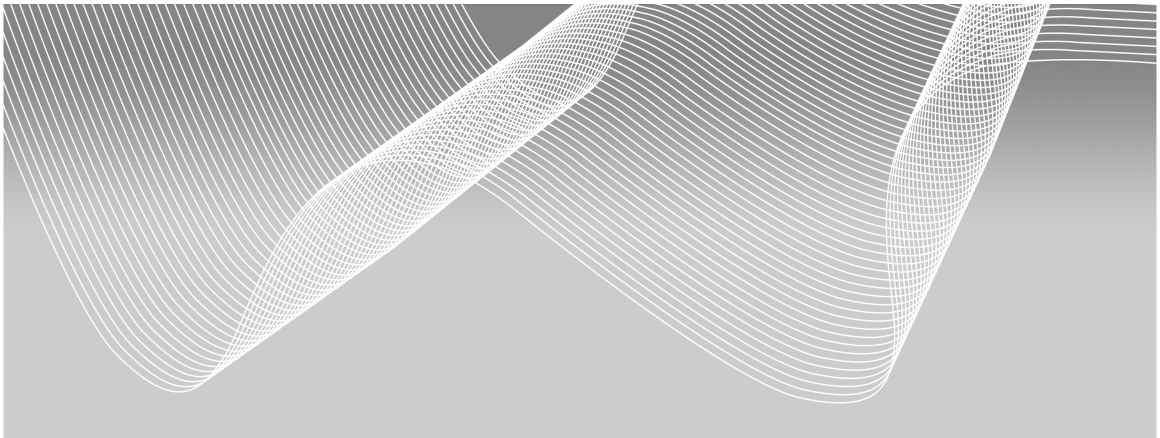


Chapter 1

Getting Started: Introduction to JMP

Goals of Data Analysis: Description and Inference

The central goal of this book is to help you build your capacity as a statistical thinker through progressive experience with the techniques and approaches of data analysis, specifically by using the features of JMP. As such, we'll begin with some remarks about activities that require data analysis, and then we'll begin using JMP.

People gather and analyze data for many different reasons. Engineers test materials or new designs to determine their utility or safety. Coaches and owners of professional sports teams track their players' performance in different situations to structure rosters and negotiate salary offers. Chemists and medical researchers conduct clinical trials to investigate the safety and efficacy of new treatments. Demographers describe the characteristics of populations and market segments. Investment analysts study recent market data to fine tune investment portfolios. All of the individuals who are engaged in these activities have consequential, pressing needs for information, and they turn to the techniques of statistics to meet those needs.

There are two basic types of statistical analysis: description and inference. We do descriptive analysis in order to summarize or describe an ongoing process or the current state of a population—a group of individuals or items that is of interest to us. Sometimes we can collect data from every individual in a population (every professional athlete in a sport, or every firm in which we currently own stock), but more often we are dealing with a subset of a population—that is to say with a sample from the population.

If a company reviews the records of all of its client firms to summarize last month's sales to all customers, the summary will describe the population of customers. If the same company wants to use that summary information to make a forecast of sales for next month, the company needs to engage in inference. When we use available data to make a conclusion about something we cannot observe, or about something that hasn't happened yet, we are drawing an inference. As we'll come to understand, inferential thinking requires risk-taking, and it can be done well or poorly. Learning to minimize the risks inherent in inference is a large part of the study of statistics.

Types of Data

The practice of statistical analysis requires data—when we "do" analysis, we're analyzing data. It's important to understand that analysis is just one phase in a statistical study.

Later in this chapter we'll look at some data collected in 1879 by Albert A. Michelson, who was measuring the speed of light. He carefully designed his experiments and his instruments, taking repeated measurements to come up with an accurate estimate of the speed of light. He did this well; in 1907 he received the Nobel Prize in Physics for this work. In the data set that we'll analyze, his measurements are grouped into a sequence of five "trials," each consisting of 20 measurements.

In this particular example we have two variables that we'll represent as two columns within a data table. A variable is an attribute that we can count, measure, or record. The two variables in the Michelson data are "Velocity" (the measurement that he wrote down) and "Trial#", indicating the group to which each individual measurement belonged. Typically, we'll record or capture multiple observations of each variable—whether we're taking repeated measurements as Michelson did, or whether we're recording facts from numerous respondents in a survey or individual items on an assembly line. Each observation (often called a case or subject in survey data) occupies a row in a data table.

Whenever we analyze a data set in JMP, we'll work with a *data table*. The *columns* of the table contain different variables, and the *rows* of the table contain observations of each variable. In your statistics course, you'll probably use the terms data set, variable, and observation (or case). In JMP we more commonly speak of data tables, columns, and rows.

One of the organizing principles you'll notice in this software is the differentiation among *data types* and *modeling types.* The columns that you will work with in this book are all either *numeric* or *character* data types, much like data in a spreadsheet are numeric or labels.

In your statistics course you may be learning about the distinctions among different kinds of quantitative and qualitative data. In JMP these distinctions are called *modeling types* and JMP recognizes three such types:

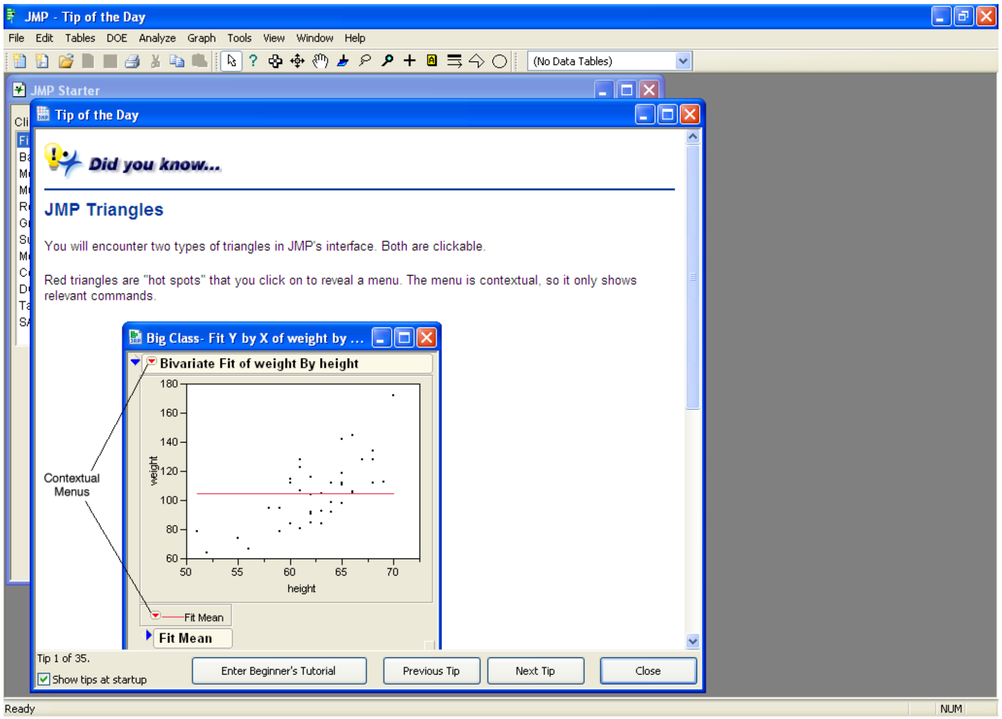
- *Continuous* columns are inherently numeric (their data type is numeric; you can meaningfully compute sums, averages, and so on), and can assume an infinite number of values. Most measurements and financial figures are continuous data. Michelson's measurements of light's velocity are continuous.
- *Ordinal* columns establish or reflect a sequence of groupings (for example, small, medium, large), chronology (for example, pre- and post-event), or any other classification with an inherent ordering of observations. In our data table, we have an ordinal variable indicating the sequence of Michelson's five measurement groups. Ordinal columns can be either numeric or character data.
- *Nominal* columns simply differentiate among groups within the data. For example, if we are analyzing health data from different countries, we might want to compare figures by continent. In that case, continent would be considered a nominal (also known as categorical) variable. Nominal variables can also be numeric or character. So names are nominal, as are postal codes or telephone numbers.

As we'll soon see, understanding the differences among these modeling types is helpful in understanding how JMP treats our data and presents us with choices.

Starting JMP

Whether you are using a Windows-based computer or a Macintosh, JMP works in very similar ways. All of the illustrations in this book were generated in a Windows environment. Find JMP[1] among your programs and launch it. You'll see the opening screen shown in Figure 1.1. The software opens a Tip of the Day window each time you start the software. These are informative and helpful. You can elect to turn off the automatic messages by clearing the **Show tips at startup** check box in the lower-left part of the window. You'll be well advised to click the **Enter Beginner's Tutorial** button sooner rather than later to get a helpful introduction to the program (perhaps you should do so now or after reading this chapter). After you've read the tip of the day, click **Close**.

Figure 1.1 The JMP Opening Screen



JMP Triangles

You will encounter two types of triangles in JMP's interface. Both are clickable.

Red triangles are "hot spots" that you click on to reveal a menu. The menu is contextual, so it only shows relevant commands.

Tip 1 of 35.

☑ Show tips at startup

[1] JMP 8.02

The next window displayed is called the JMP Starter window, which is an annotated menu of major functions. It is worth your time to explore the JMP Starter window by navigating through its various choices to get a feel for the wide scope of capabilities that the software offers. As a new user, though, you may find the range of choices to be overwhelming.

In this book, we'll tend to close the JMP Starter window and use the menu bar at the top of the screen to make selections.

A Simple Data Table

In this book, we'll most often work with data that has already been entered and stored in a file, much like you would type and store a paper in a word-processing file or data in a spreadsheet file. In Chapter 2, you'll see how to create a data table on your own.

We'll start with the Michelson data mentioned earlier.

1. Click **File → Open**.
2. Navigate your way to the folder of data tables that accompany this book.[2]
3. Select the file called **Michelson 1879** and click **Open**.

The data table appears in Figure 1.2. Notice that there are four regions in this window: three vertically arranged panels on the left, and the data grid on the right.

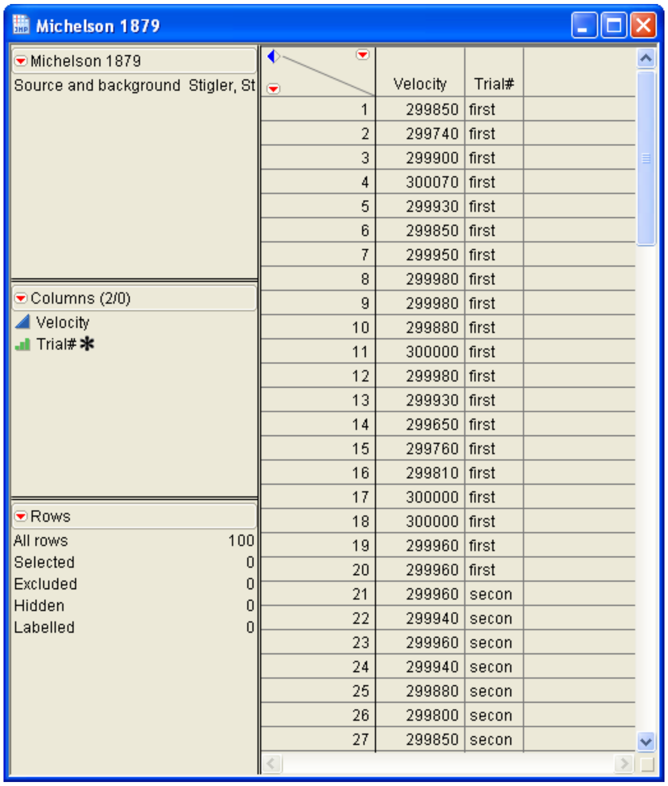
The three panels provide *metadata* (descriptive information about the data in the table). In addition to displaying the metadata, the entries in the panel are editable, so you can change metadata. Later we'll discuss these panels in greater detail. For now, let's just get oriented.

Beginning at the top left, we find the *Table* panel, which displays the name of the data table file as well as optional information provided by the creator of the table. You'll see a small red triangle pointing downward next to the table name.

[2] All of the data tables used in this book are available from support.sas.com/authors. If you are enrolled in a college or university course, your instructor may have posted the files in a special directory. Check with your instructor.

Red triangles indicate a context-sensitive menu, and they are an important element in JMP. We'll discuss them more extensively a few pages from now, but you should expect to make frequent use of these little red triangles.

Figure 1.2 The Michelson 1879 Data Table



Michelson 1879

Michelson 1879
Source and background Stigler, St

Columns (2/0)
Velocity
Trial# ✱

Rows

All rows	100
Selected	0
Excluded	0
Hidden	0
Labelled	0

	Velocity	Trial#
1	299850	first
2	299740	first
3	299900	first
4	300070	first
5	299930	first
6	299850	first
7	299950	first
8	299980	first
9	299980	first
10	299880	first
11	300000	first
12	299980	first
13	299930	first
14	299650	first
15	299760	first
16	299810	first
17	300000	first
18	300000	first
19	299960	first
20	299960	first
21	299960	secon
22	299940	secon
23	299960	secon
24	299940	secon
25	299880	secon
26	299800	secon
27	299850	secon

Just below the red triangle, there is a note describing the data and identifying its source. You can open that note (called a *Table variable*) just by double-clicking on the words beginning "Source and background." Figure 1.3 shows what you'll see when you double-click. A table variable contains metadata about the entire table.

Figure 1.3 Table Variable Dialog Box

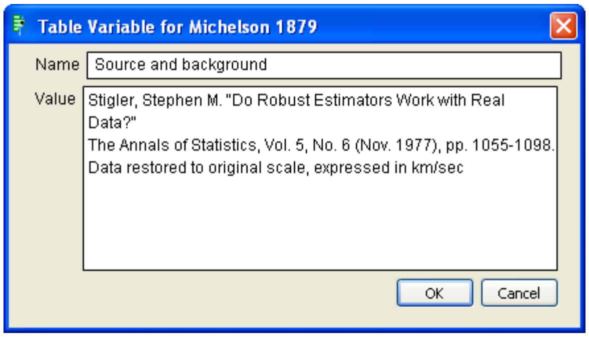


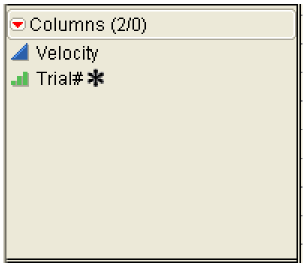
Table Variable for Michelson 1879

Name: Source and background

Value: Stigler, Stephen M. "Do Robust Estimators Work with Real Data?"
The Annals of Statistics, Vol. 5, No. 6 (Nov. 1977), pp. 1055-1098.
Data restored to original scale, expressed in km/sec

Below the **Table** panel is the *Columns* panel, which lists the column names and JMP modeling types, as well as other information about the columns. This will be explained further below, but for now, let's take note of a few important landmarks and concepts.

Figure 1.4 Columns Panel



Columns (2/0)
Velocity
Trial# ✱

The notation **(2/0)** in the top box of the panel tells us that there are two columns in this data table, and that neither of them is *selected* at the moment. In a JMP data table, we can select one or more columns or rows for special treatment. There is much more to learn about the idea of selection, and we'll return to it later in this chapter.

Next we see the names of the two columns. To the left of the names are icons indicating the modeling type. In this example, the blue triangle next to **Velocity** means that the column contains a continuous variable.

The green ascending bar icon next to **Trial#** indicates an ordinal variable. This data table doesn't include a nominal variable, but the corresponding icon looks like a red bar graph.

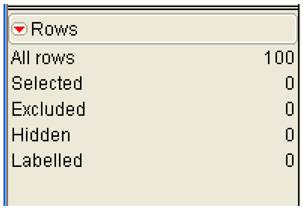
You'll also notice an asterisk next to the **Trial#** variable. This is one of several special symbols, which, in this case, indicates that the creator of the data table specified the order of values for this variable (first, second, third, and so on). If no order had been specified, JMP would sequence the values alphabetically.

Finally, we find the *Rows* panel (Figure 1.5), which provides basic information about the number of rows (in this case 100). Like the other two panels, this one provides quick reference information about the number of rows and their *states*.

> The idea of “row states” is an important one in JMP, and probably will seem unfamiliar. It is fairly simple and quite useful, and is discussed more fully later in the chapter in the “Row States” section.

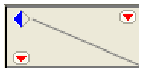
The top row indicates that there are 100 observations in this data table. The next four rows refer to the four basic row states in a JMP data table. Initially, all rows share the same state, in that none has been selected, excluded, hidden or labeled. Row states enable us to control whether particular observations appear in graphs, are incorporated into calculations, or whether they are highlighted in various ways.

Figure 1.5 Rows Panel



Rows	
All rows	100
Selected	0
Excluded	0
Hidden	0
Labelled	0

The *Data Grid* area of the data table is where the data reside. It looks like a familiar spreadsheet format, and it just contains the raw data for any analysis. Unlike a spreadsheet, each cell in the data grid contains a data value, but never a formula. The data values might be the result from a computation, but we cannot place a formula directly into a cell. We can assign a formula to an entire column, but not to one cell.



In the upper-left corner of the data grid, you’ll see the region shown here. There is a diamond-shaped *disclosure button* (with blue shading on the left side here in Windows; on a Macintosh it is an arrowhead ▶). Disclosure buttons allow you to expand or contract the amount of information displayed on the screen. The disclosure button show here lets you temporarily hide the three panels discussed above.

4. Try it out! Click on the disclosure button to hide and then reveal the panels.

The red triangles offer you menu alternatives that won’t mean much at this point, but which we’ll discuss in the next section. The hotspot in the upper-right corner (above the diagonal line) refers to the columns of the grid, and the one in the lower-left corner to the rows.

The very top row of the grid contains the column names, and the left-most column contains row numbers. The cells contain the data.

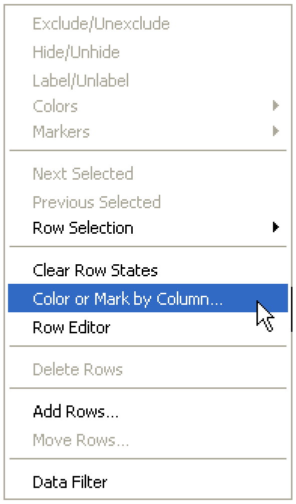
Hot Spots

We have seen several red triangular hot spot icons thus far, and it's time to take a closer look. We'll gradually learn more about using the menus they open; for now, let's look at one example of their usefulness.

1. Before going further, enlarge the data grid by clicking and dragging the lower right-hand corner.

2. In the upper-left portion of the data grid, click on the rows hotspot. When you do so, you'll see the list of menu choices shown in Figure 1.6.

Figure 1.6 Rows Hot Spot Menu



Some of the options appear faintly in gray; these choices are not available. Generally speaking, the available options enable us to select rows for special treatment, to edit the content of rows, to add rows, or to customize the appearance of rows in the table. In this first chapter, let's color code the observations in each of the five trials.

3. In the menu, select **Color or Mark by Column**.

This opens the dialog box in Figure 1.7.

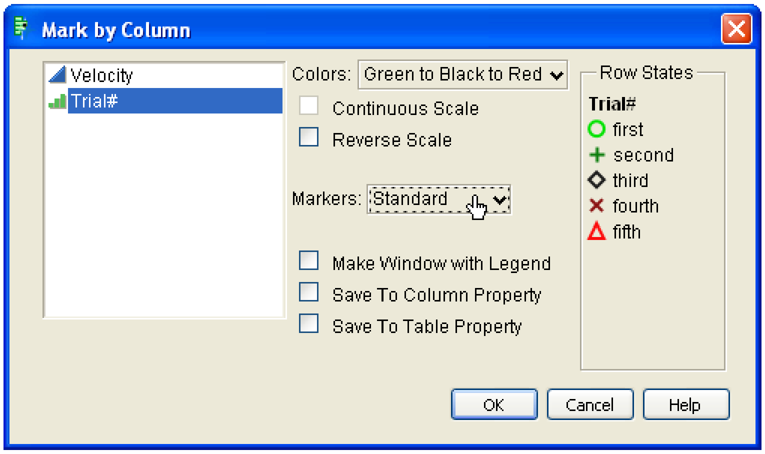
4. We want to base our color coding on the five different values of the **Trial#** variable, so we select that variable as shown in the figure.

5. Select the **Green to Black to Red** color scheme.

6. Select the **Standard** marker set, and then click **OK**.

Now look at the data grid, and scroll through the rows. You'll see that all of the first-group measurements now display a green circle next to the row number. The second group is marked with a dark green plus sign, and so on. Later, when we create a graph, these color-coded symbols will provide a quick visual reference for the different groups.

Figure 1.7 Mark by Column Dialog Box



Mark by Column

Velocity
Trial#

Colors: Green to Black to Red

□ Continuous Scale
□ Reverse Scale

Markers: Standard

□ Make Window with Legend
□ Save To Column Property
□ Save To Table Property

Row States
Trial#
○ first
+ second
◇ third
× fourth
△ fifth

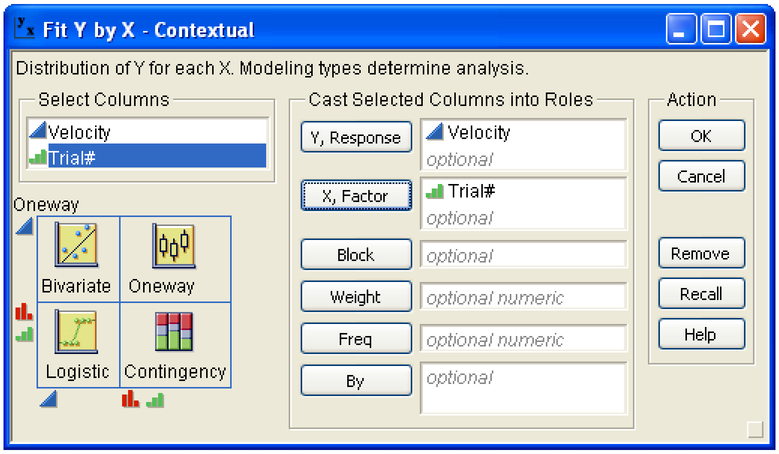
Analysis Platforms—A First Descriptive Analysis

We have Michelson's data, so let's take a look at his measurements. As you peruse the list of values, you might notice that they vary. Variation is so typical as to be almost invisible, but the very fact that they vary is what leads us to analyze them. Think about it. Michelson was measuring something that we now think of as a constant: the speed of light. In his day, the idea that light travels at a constant speed was not universally accepted, and what's more he and his contemporaries did not know the value of that constant. Hence, he was trying to measure this unknown constant. His instrumentation was imperfect, and he couldn't look up the correct value in the back of his physics textbook.

Let's see what he was coming up with in 1879. We have a table displaying all 100 measurements; now let's make a simple graph to summarize the table information visually, displaying the 20 measurements in each trial group. Don't worry about the details of these steps. The goal right now is just for you to see a typical JMP platform and its output.

1. Select **Analyze → Fit Y by X**. This *analysis platform* lets us plot one variable (the speed of light measurements) versus another (trial group).
2. In this dialog box (Figure 1.8), we'll cast **Velocity** as the **Y** or **Response** variable[3] and **Trial#** as the **X** variable, or **factor**. Click **OK**.

Figure 1.8 Fit Y by X Dialog Box



Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

Select Columns: Velocity, Trial#

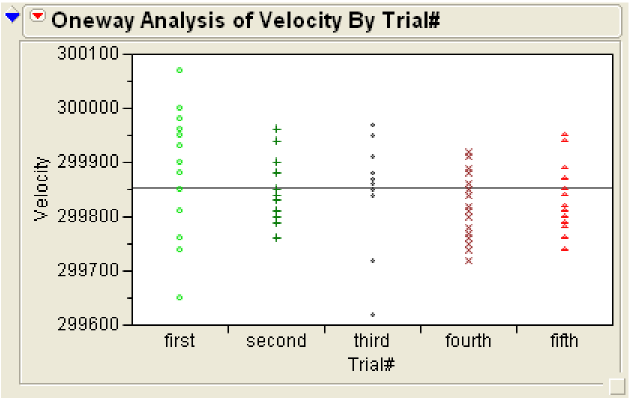
Cast Selected Columns into Roles

Y, Response	Velocity optional
X, Factor	Trial# optional
Block	optional
Weight	optional numeric
Freq	optional numeric
By	optional

You'll see this graph (Figure 1.9) as a result:

[3] In Chapter 4 we will study *response variables* and *factors*. In this chapter we are getting a first look at how analysis platforms operate.

Figure 1.9 Oneway Analysis of Velocity By Trial#



In the graph, we see the trials listed sequentially on the horizontal axis and velocity values on the vertical. Each observation is marked with a colored symbol representing the measurement of light speed. The horizontal line at approximately 299,850 km per second is the average (the mean) of all 100 observations. From the graph, we can readily see that the measurements in the first trial group were far less consistent than in the second trial, but that the second group of measurements seems to have been fairly balanced above and below the average. The fourth set of measurements was comparable to the second in terms of consistency, but tends to be lower than the measurements in the second group.

If you look closely at the graph you may realize that there aren't 20 symbols for each trial. That is because some values are so close together that the markers overlap. We can tweak the graph to plot a separate marker for each point:

3. Click on the red triangular hot spot, and choose **Display Options**. Choose **Points Jittered**, which moves the individual points ever so slightly. Notice how the graph changes.

4. Click on the hot spot again, choose **Display Options** and choose **Connect Means**.

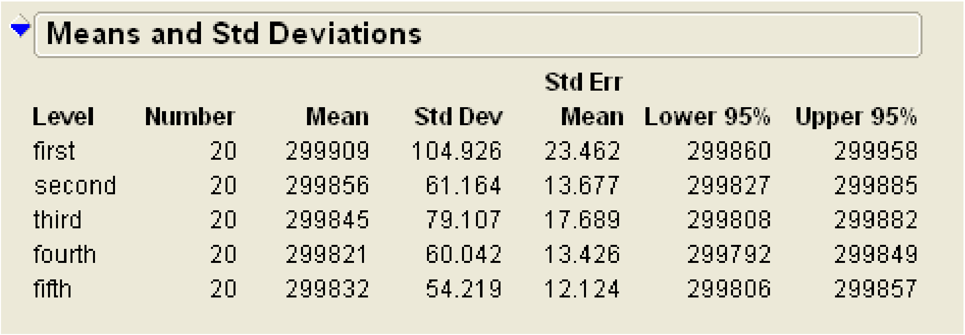
Look again at the modified graph. The new blue line on your graph represents the mean of the 20 measurements in each trial. What was happening to Michelson's measurements as he repeated these trials?

Finally, the graph now shows us the mean values of each group. Suppose we want to know the numerical values of the five averages.

5. Click the hot spot once more, and this time choose **Means and Std Dev** (standard deviations).

This will generate a table of values beneath the graph, as shown in Figure 1.10. For the current discussion, we'll focus our attention only on the first three columns. Later in the book we'll learn the meaning of the other columns. This table (below) reports the mean for each of the five trial groups, and also reports that there are 20 observations in each group.

Figure 1.10 Table of Means and Standard Deviations



Means and Std Deviations

				Std Err		
Level	Number	Mean	Std Dev	Mean	Lower 95%	Upper 95%
first	20	299909	104.926	23.462	299860	299958
second	20	299856	61.164	13.677	299827	299885
third	20	299845	79.107	17.689	299808	299882
fourth	20	299821	60.042	13.426	299792	299849
fifth	20	299832	54.219	12.124	299806	299857

Row States

Our data table consists of 200 cells: two variables with 100 observations each, arrayed in two columns and 100 rows. One guiding principle in statistical analysis is that we generally want to make maximum use of our data. We don't casually discard or omit any portion of the data we've collected (often at substantial effort or expense). There are times, however, that we might want to focus attention on a portion of the data table or examine the impact of a small number of extraordinary observations.

By default, when we analyze one or more variables using JMP, every observation is included in the resulting graphs and computations. You can use **row states** to confine the analysis to particular observations or to highlight certain observations in graphs.

There are four basic row states in JMP. Rows can be one of the following:

- *Selected*: selected rows appear bolded or otherwise highlighted in a graph.
- *Excluded*: when you exclude rows, those observations are temporarily omitted from calculated statistics such as the mean. The rows remain in the data table, but as long as they are excluded they play no role in any computations.
- *Hidden*: when you hide rows, those observations do not appear in graphs, but are included in any calculations such as the mean.

- *Labeled*: The row numbers[4] of any labeled rows display next to data points in some graphs for easily identifying specific points.

Let's see how the row states change the output that we've already run by altering the row states of rows 3 and 4.

1. First, arrange the open windows so that you can clearly see both the **Fit Y by X** window and the data table.
2. Move your cursor into the data table, and select rows 3 and 4 by clicking and dragging on the row numbers 3 and 4. You'll see the two rows highlighted within the data table.

Look at your graph. You should see that two of the green circles among the first trial data are larger than the others. That's the effect of selecting these rows. Notice also that the **Rows** panel now shows that two rows have been selected.

3. Click on another row, and then drag your mouse slowly down the column of row numbers. Do you notice the rows highlighted in the table and the corresponding symbols "lighting up" in the graph?
4. Press Esc or click in the triangular area above the row numbers to deselect all rows.

Next we will exclude two observations and show that the calculated statistics change when they are omitted from the computations. To see the effect, we first need to instruct JMP to automatically recalculate statistics when the data table changes.

5. Click the red triangle next to **Oneway Analysis** in the report window and choose **Script → Automatic Recalc**.
6. Now let's exclude rows 3 and 4 from the calculations. To do this, first select them as you did before.
7. Select **Rows → Exclude/Unexclude** (in Windows, you can also find this choice by right-clicking). This will exclude the rows.

[4] Columns can contain labels (for example, the name of respondent or country name) which are also displayed when a row is labeled.

Now look at the analysis output. The number of observations in the first group is now 18 rather than 20 and the mean value for that group has changed very slightly. Toggle between the exclude and unexclude states of these two rows until you understand clearly what happens when you exclude observations.

8. Finally, let’s hide the rows. First, be sure to unexclude rows 3 and 4 so that all 100 points appear in the graph and in the calculations.

9. With rows 3 and 4 still selected, **Rows → Hide/Unhide.** This will hide the rows (check out the very cool dark glasses icon).

Look closely at the graph and at the table of means. The two enlarged green circles are missing, but there are still 20 observations in the first trial group.

Exporting JMP Results to a Word-Processor Document

As a statistics student you may often want or need to include some of your results within a paper or project that you’re writing for class. As we wrap up this first lesson, here’s a quick way to capture output and transfer it to your paper. To follow along, first open your word processing software, and write a sentence introducing the graph you’ve been working with. Then return to the JMP analysis output window.

Our analysis includes a graph and a table. To include the graph only, do this:

1. Select **Tools → Selection**. Your cursor will now become an open cross.

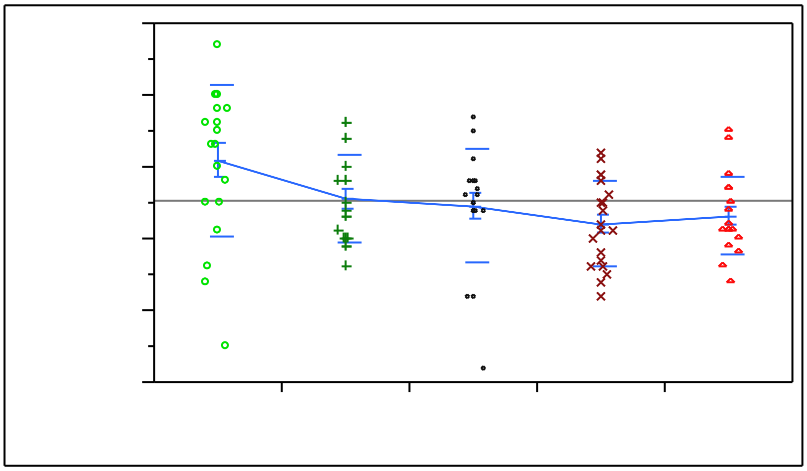
2. Click and drag the cursor across the graph until the entire graph is highlighted.

3. Select **Edit → Copy.**

4. Now move to your word processor and paste your copied graph.

The graph should look something like the one shown below in Figure 1.11. Note that the graph will look slightly different from its appearance within JMP, but this demonstration should illustrate how very easy it is to incorporate JMP results into a document.

Figure 1.11 A Graph Pasted from JMP



Saving Your Work

As you work with any software you should get in the habit of saving your work as you go. JMP supports several types of files, and enables you to save different portions of a session along the way. You've already seen that data tables are files; we've modified the **Michelson 1879** data table and might want to save it.

Alternatively, you can save the *session script*, which essentially is a transcript of the session—all of the commands you issued, as well as their results. Later, when you restart JMP, you can open the script file, run it, and your screen will be exactly as you left it.

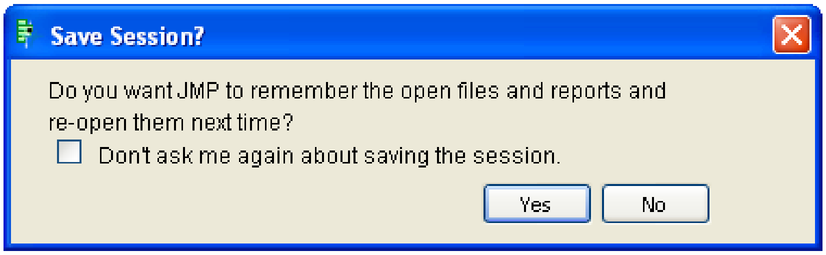
5. Select **File → Save Session Script**. In the dialog box, choose a directory in which to save this JSL file, give the file a name, and click **OK**.

Leaving JMP

We've covered a lot of ground in this first session, and it's time to quit.

1. Select **File → Exit**.

Answer **No** to the question about saving changes to the Michelson data. Then you'll see this dialog box:



Do you want JMP to remember the open files and reports and re-open them next time?

□ Don't ask me again about saving the session.

In this case, you can click **No**. In future work, if you want to take a break and resume later where you left off, you may want to click **Yes**. The next time you start the program, everything will look as it did when you quit.

Remember to run the Beginner's Tutorial before moving on to Chapter 2.